Sentiment Analysis of Polish Texts

Kamila Kowalska, Di Cai, and Steve Wade

***Abstract*—A new language resource for sentiment analysis (SA) and an application of SA to a new domain—discussions on an online Polish news forum—are developed. A scheme for human annotation of textual samples is proposed using online questionnaires. A method for classifying the samples based on the annotations is introduced and put into practice. A method, applying existing advanced Information Retrieval (IR) techniques, for SA within a Bayesian learning framework is explored. Preliminary experimental results show the IR techniques, in conjunction with the Naïve Bayes classifier, can be expected to produce good sentiment classification performance both for Polish texts and for the news discussion domain.**

***Index Terms*—Polish texts, web data creation, human annotation, machine learning, sentiment analysis.**

I. Introduction

The majority of studies on sentiment analysis (SA) deal with textual content written in English. Work in this area is focussed on the use of computational techniques to extract subjective information or, at least, to identify the polarity of opinions expressed in source texts. Given the variety of techniques and algorithms developed for English, it is an important issue to test and verify their performance for other languages. Recently this has been thoroughly studied for Chinese [9]. In this study we focus on SA using texts written in Polish, which is a research area still quite unexplored. To our best knowledge, there has been only one previous study related to SA, which referred to the automated classification of Polish product reviews [1].

Both Polish and English belong to the great family of Indo-European languages that contains several hundred languages and dialects spoken nowadays by 45% of the global population. They are, however, members of different subgroups with distinct history and morphological evolution over the centuries: English belongs to the Germanic family, while Polish to Balto-Slavic. This results in many linguistic differences between the two languages. First of all, the structure of a single Polish sentence, namely, the order of subsequent parts of speech, is not so strongly constrained as in English. Nouns, verbs and adjectives may appear at the beginning, in the middle, or at the end of the sentence without changing its meaning. That could make the task of automatic identification of the parts of speech somehow tricky. What is more, personal pronouns are very often omitted and the subject or object of the sentence can be identified only by inflection. Secondly, Polish is a highly inflected language with very rich and complex morphology. Nouns, pronouns, adjectives and numerals come under declension, indicating such features as gender (masculine, feminine, neuter), case and number (singular and plural). There are several schemes of declensions, as well as many irregularities and exceptions from the main rules. Verbs inflect according to gender, number, tense, mood and voice. Most of them occur in two aspects: imperfective and perfective.

The above statements make the case for the difficulty of transferring linguistic techniques, developed for English, to Polish for SA. They also make the point that statistical techniques may be inherently more promising for multi-language application. For many statistical methods, each document is represented as a vector of a set of terms, without considering word orders, grammars, sentence structures and the roles of the parts of speech. Therefore, this study attempts to explore a method applying SA techniques, originally developed for English texts, to texts written in Polish in a Bayesian learning framework. The Naive Bayes classifier (NBC) is surprisingly effective in practice since its classification decision can be correct even if its probability estimates are inaccurate [5,6]. There are theoretical reasons [10] for NBC's apparently unreasonable effectiveness. An advantage of NBC is that it requires a small amount of training data to estimate the parameters necessary for classification. Also, the independence assumption enables the parameters to be learned separately and thus greatly simplifies learning processing, especially when the number of terms used to index documents is extremely large.

The paper is organized as follows. Web data creation is developed in Section 2. A scheme for human annotation of samples is proposed and an annotation method for classifying the samples is introduced in Section 3. A machine learning method for SA applying existing advanced IR techniques is explored and experimental results are shown in Section 4. Conclusions are drawn in Section 5.

II. Web Data Creation

News forums are web sites which offer a constant flow of hottest news published on their front pages, as well as an opportunity for readers to express their opinions and sentiments on a great variety of topics. Playing the role of on-line and national newspapers, they gather a significant number of users representing a broad spectrum of community. It is important that people do not limit their forum activity to commenting on the main topic only. Discussion development often enhances the level of extreme opinions and sentiment. For these reasons News forums seem to be a very good experimental environment for

Manuscript received February 25, 2012; revised April 25, 2012.

K. Kowalska is with National Center for Nuclear Research, Świerk, Poland.

D. Cai and S. Wade is with School of Computing and Engineering, University of Huddersfield, HD1 3DH, UK (e-mail: d.cai@hud.ac.uk)

studying SA.

Data was created for our study from the *TVN24* News forum (http://www.tvn24.pl/forum.html). We crawled the discussion regarding five most active topics on the 11th of March 2009. The whole dataset (collection) consists of 1000 texts (documents) with length longer than 10 and shorter than 100 distinct words. The reason for introducing the length restriction is that short texts, though they could express a very distinct sentiment, do not provide enough input for the classification algorithms. On the other hand, the probability of coexistence of both positive and negative opinions, often on different subjects, tends to be higher in longer texts than in short ones. Since the SA techniques may be expected to perform better if each text contains only a single opinion, it seems reasonable to reduce the number of multi-opinion texts by limiting the text length.

a	aby	acz	aczkolwiek	ale	ależ	aż
bardziej	bardzo	bez	bo	bowiem	by	byli
być	był	była	było	były	będzie	będą
cali	cała	cały	choć	chociaż	co	cokolwiek
coś	czasami	czasem	czemu	czy	czyli	...

Fig. 1. A list of stop words

The dataset was transformed in order to reduce the level of noise. The transformations included: case flattening, numbers removal, stop words removal and stemming. Stopwords for Polish were chosen as the most common pronouns, prepositions, conjunctives and particles. Specially, we added also two verbs ('to be' and 'to have') that are very often used but do not carry any sentiment. The complete list of stop words has 207 words. Some of them are shown in Fig. 1. Stemmer for Polish was based on a freeware morphology dictionary downloaded from http://morfologik.blogspot.com/. The important feature of this tool is, if a word stemmed does not exist in the dictionary, it is simply removed from the transformed document. This often removes misspellings. The average quantity of the removed content is about 20% ± 0.7. The distribution of the fraction of the removed content is presented in Fig. 2. No document was reduced by more than 50% after removal and stemming. Although the transformation may cause information loss, the resulting dataset is much cleaner.

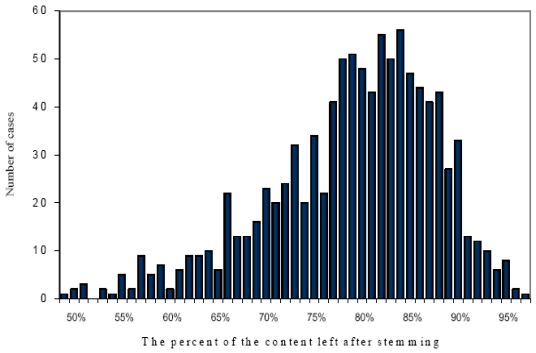

Fig. 2. Distribution of the fraction of the textual content left after stemming.

III. Human Annotation Method

The human annotators were chosen from a group, so-called 'panel', of people who agreed to fill in on-line GEMIUS questionnaires. At the beginning, the panel consisted of about 32,000 people (panelists) and its demographic structure reflected, to some extent, the structure of the whole Polish community. Each panelist was asked to read texts and then verify whether the emotional content of each text was positive, negative or neutral. There was also the fourth answer 'no idea' for those who felt unable to categorize the sentiment of the text. The questionnaires were randomly sent to 12,221 people, 2,214 of them (called "annotators") in the end decided to fill it in completely. Each annotator was asked to classify ten different, randomly chosen texts. Each text was classified by 30 different annotators. After the necessary amount of questionnaires had been gathered, all answers 'no idea' were discarded, and the qualitative scale of the questionnaire was then transformed to a numeric scale called sentiment score. The final emotional value of each text was thus based on only three answers, they were derived from the average of the thirty different sentiment scores.

More specifically, let D be the collection of documents and $A = \{a_1, a_2, \dots, a_r\}$ be a group of r annotators. Let $a_i(d)$ be a sentiment score that annotator a_i assigns to document $d \in D$. Then the qualitative scale of the questionnaire may be transformed to the numeric scale by assigning the answers from the following score function:

$$a_i(d) = \begin{cases} 1 & d \text{ is negative} \\ 2 & d \text{ is neutral} \\ 3 & d \text{ is positive} \end{cases}$$

where $i = 1, 2, \dots, r$. Then we may classify document d by the following classification *rule*-1:

$$cla_1(d) = \begin{cases} negative\ class & \text{the number of people assigning } a(d) \text{ to } 1 \geq 50\% \\ neutral\ class & \text{the number of people assigning } a(d) \text{ to } 2 \geq 50\% \\ positive\ class & \text{the number of people assigning } a(d) \text{ to } 3 \geq 50\% \end{cases}$$

For instance, if more than 50% of the scores (out of r scores) are 3, then d is classified as being positive.

For the case where the *rule*-1 is not satisfied, the average of r different scores is considered for assigning the final emotional value. Let us denote

$$emv(d) = \frac{1}{r} \sum_{a_i \in A} a_i(d)$$

The average may be used to determine sentiment orientation of each document. That is, we classify each document with the following classification *rule*-2:

$$cla_2(d) = \begin{cases} negative\ class & 1.0 \leq emv(d) \leq 1.7 \\ neutral\ class & 1.7 \leq emv(d) < 2.3 \\ positive\ class & 2.3 \leq emv(d) \leq 3.0 \end{cases}$$

The final classification results, for our dataset crawled from the *TVN24* news forum (with $r = 30$) were: 668 negative documents, 132 positive documents and 200 neutral. The annotated samples will appear in our experiments at a later stage.

There are several statistics that measure the reliability of agreement achieved between more than two different annotators coding the same documents. Fleiss' κ [2] measures the consistency of the ratings. It assumes that the different items are rated by different annotators, although the total number of the annotators per item is fixed. It is defined as:

$$\kappa = \frac{p_o - p_c}{1 - p_c}$$

where p_o describes the agreement observed in the experiment and p_c denotes chance agreement, that is, the probability that independent annotators would agree if they rated the item choosing the answers randomly. $\kappa = 1$ is for perfect agreement between annotators; $\kappa \leq 0$ is for when annotators disagree to a larger extent than would occur by chance. An alternative reliability coefficient is given by Krippendorff's α [3]:

$$\alpha = 1 - \frac{q_o}{q_e}$$

where q_o denotes the observed disagreement between the annotators and q_e is disagreement that would be observed for independent annotators rating the items randomly. We calculated both statistics for our dataset and obtained: $\kappa = 0.1815$ and $\alpha = 0.1457$. The level of agreement between the annotators seems not impressive, but one could have expected such a result, given the number of different annotators classifying a single document.

IV. Machine Learning Method

The Naive Bayes (NB) classifier is used in this study. Let $X \subseteq D$ be a sentiment class (positive, negative, or neutral). For a possible class X, this model computes the posterior probability, $p(X|d)$, that document d belongs to X. Then it classifies d into the class with the highest posterior probability.

More specifically, let $V=\{t_1, t_2, \ldots, t_n\}$ be a vocabulary of terms used to index individual documents in D. With the conditional independence assumption that the presence (or absence) of a term is independent of the presence (or absence) of other terms, we can write

$$p(X|d) = \frac{p(X)}{p(d)} \cdot \prod_{i=1}^{n} p(t_i|X)\, p(X) \cdot \prod_{i=1}^{n} p(t_i|X)$$

where $p(t_i|X)$ is the conditional probability of term t_i occurring in some document of class X, $p(d)$ is the probability that a randomly picked document is d, and $p(X)$ is the probability that a randomly picked document belongs to class X. The parameters, such as, *a priori* probability $p(X)$ and the posterior probability $p(t|X)$ $(t \in V)$ can be estimated by the following formula.

$$p(X) = \frac{|X|}{|D|} \quad \text{and} \quad p(t|X) = \frac{\sum_{d \in X} w_d(t)}{\sum_{d \in X} \sum_{t' \in V} w_d(t)}$$

where $w_d(t)$ is a *weighting function* estimating the importance of term t in representing d.

Four weighting functions were used in our experiments: The inverse document frequency of term t in X, $idf_X(t) = \log \frac{n_X(t)}{|X|}$ where $n_X(t)$ is the number of documents of class X in which t occurs. The second is

$$\Delta idf = idf_X(t) - idf_{\overline{X}}(t) = \log \frac{p(X)}{1 - p(X)} - \log \frac{n_X(t)}{n_{\overline{X}}(t)}$$

Study [4] showed good performance using Δidf, along with SV, for sentiment classification. The third was the *Okapi* weighting function [7]:

$$w_d(t) = \frac{(a+1) \cdot f_d(t)}{a \cdot \left[(1-b) + b \cdot \frac{L_d}{ave(X)}\right] + f_d(t)}$$

where parameters $a = 1.2$ and $b = 0.75$, $f_d(t)$ is the frequency of term t in document d, L_d is the length of document d and $ave(X)$ is the average length of documents in X. The last one is the *Smart* weighting function [8]:

$$w_d(t) = \frac{\left[1 + \ln\left(1 + (\ln f_d(t))\right)\right] \cdot \log\left(\frac{|X|+1}{\mathrm{L_d}}\right)}{(1-c) + c \cdot \frac{L_d}{ave(X)}}$$

where parameter $c = 0.2$. Both the *Okapi* and *Smart* functions have widely been recognized to produceexcellent retrieval performances in IR.

Preliminary experiments were carried out using the samples classified by the annotation method as described in Section 3. 4-fold cross-validation and the standard measures recall (the proportion of correct documents actually classified) and precision (the proportion of classified documents actually correct) were used for evaluation. The experimental results are displayed in Table I.

TABLE I: Performance Comparisons with Three $w_d(t)$

	Negative Classes		Positive Classes	
$w_d(t)$	Recall	Precision	Recall	Precision
tf	0.8870	0.9120	0.8967	0.9130
Δidf	0.9110	0.9294	0.9188	0.9290
Okapi	0.9010	0.9142	0.9040	0.9169
Smart	0.9032	0.9156	0.9010	0.9147

From the results it can be seen: (1) classifications obtained using the four weighting functions achieved good performances (above 90% recall/precision) at most evaluation points; (2) *Okapi* and *Smart* showed similar performances at all the evaluation points; (3) $tf * idf$ achieved consistently better performances than *tf*, *Okapi* or *Smart*; the improvements were shown at all the evaluation points; (4) *tf*, the simplest weighting function, obtained comparable performance with sophisticated weighting functions in our current experiments; this confirmed some past experimental studies emphasising that document representation with term frequency weighting can produce good performance for sentiment classification tasks.

V. Conclusions

In this study, we discussed the techniques of data creation

and proposed a scheme for human annotation of textual samples using online questionnaires. The values of standard reliability measures, Fleiss' κ and Krippendorff's α, showed slight agreement between thirty annotators labelling each sample. We then introduced a method to classify the samples based upon the annotations. We next investigated the performance of sentiment classification using the NB method against the classification obtained from the annotation method. Preliminary experimental results showed the NB method, when used in conjunction with existing advanced IR techniques of document representation, can be expected to achieve good performance, both for the news discussion domain and for Polish texts.

REFERENCES

[1] A. Buczynski and A. Wawer, "Automated classification of product review sentiments in Polish," *Akademicka Oficyna Wydawnicza*, EXIT, 2008.

[2] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76. pp. 378–382. 1971.

[3] K. Krippendorff, "Content Analysis: An Introduction to its Methodology," *Sage Publications*, Thousand Oaks, CA. 2004.

[4] J. Martineau, T. Finin, and D. tfidf, "an improved feature space for sentiment analysis," In *Proc. of the Third AAAI International Conference on Weblogs and Social Media*, 2009.

[5] A. McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," In *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[6] I. Rish, "An empirical study of the naive Bayes classifier," In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 2001.

[7] S. E. Robertson and S. Walker, "Okapi/Keenbow at TREC-8," In *the 8th Text REtrieval Conference (TREC-8)*, pp. 151–161, 1999.

[8] A. Singhal, J. Choi, D. Hindle, D. Lewis, and F. Pereira, "AT and T at TREC-7," In *the 7th Text Retrieval* Conference *(TREC-7)*, pp. 239–252, NIST Special Publication, 1999.

[9] C. Zhang, D. Zeng, J. Li, F. Wang, and W. Zuo, "Sentiment analysis of Chinese documents: From sentence to document level," *Journal of the American Society for Information Science and Technology*, vol. 60.no. 12, pp. 2474–2487. 2009.

[10] H. Zhang. The optimality of naive Bayes, in *The 17th* International *FLAIRS Conference*, 2004.